# 基于菜谱与微博用户评论的饮食社区挖掘研究*

吴小兰[1,2] 章成志[2,3]

[1](安徽财经大学管理科学与工程学院 蚌埠 233030)

[2](南京理工大学信息管理系 南京 210094)

[3](江苏省数据工程与知识服务重点实验室 南京 210093)

**摘要:** 【目的】以大规模真实社交网络数据作支撑研究饮食社区结构。【方法】使用"美食杰"网站的菜谱信息和新浪微博上与菜有关的微博数据，完成用户与菜之间的"提及"关系构建后，分别在省份地区维度和地区菜系维度进行映射，并运用社区发现算法进行社区挖掘。【结果】在省份地区关系网和地区菜系关系网上存在明显的社区结构。【局限】实验过程中发达地区人数与边缘地区人数悬殊太大，对本文所得结论有一定的影响。【结论】实证结果发现：省份地区被划分成"其他口味"、"鲜咸味"、"香辣味"三个口味地区；"川菜"、"云贵菜"因辅料独特很少与其他菜系被一起点餐，"京菜"、"沪菜"、"鲁菜"、"东北菜"常被一起点餐，除此之外，地区菜系之间存在一定程度的地理位置近邻性。

**关键词:** 饮食文化 地方菜系 饮食社区 Web 信息组织

**分类号:** G353



## 1 引 言

饮食是人类社会亘古不变的生活主题，也是人类从事其他社会活动的基础和前提。随着生产力的发展，尤其是农业的发展，中国人对于"吃"已不仅仅是果腹那么简单，围绕着"吃"产生了一系列饮食文化，如饮食风俗、饮食思想、饮食行为等。饮食文化作为与"吃"、"喝"相关的一种文化现象，它不分种族国界，与每个人息息相关，因此开展中国饮食文化研究很有必要。尽管目前关于饮食文化研究跨越不同的学科种类，研究方法也各具特色，研究成果也十分丰富，但多数研究使用的是对菜谱的定性(如对饮食文化的发生、发展脉络的梳理)和定量分析(如对关于饮食文化的文献、史料进行统计整理)方法。实际上，随着互联网和大数据技术的发展，将有利于在很大规模的真实数据集上开展饮食相关的研究，通过真实数据集的挖掘验证一些重要结论，甚至发现一些新的有实际价值的结论，因此本文开展基于菜谱与微博用户饮食评论的饮食社区挖掘研究。

## 2 相关研究概述

目前，我国研究饮食文化的群体主要是高等院校的学者、饮食行业的从业者、文学作家[1]。相对其他学科而言，学术界对于饮食文化的研究起步较晚，且近几年学术界开始着手利用真实数据来研究饮食文化。总结现有研究，笔者发现常被用来研究饮食文化的两类真实数据有：菜谱数据和用户饮食数据。

通讯作者：章成志，ORCID：0000-0001-8121-4796，E-mail：zhangcz@njust.edu.cn。

*本文系国家社会科学基金项目"在线社交网络中基于用户的知识组织模式研究"(项目编号:14BTQ033)、安徽省教育厅人文社会科学项目"基于社交网络的交叉学科知识发现及其应用研究"(项目编号:SK2016A0025)和江苏省数据工程与知识服务重点实验室开放课题"在线社交网络上交叉学科用户知识结构发现及其兴趣演变研究"(项目编号:DEKS2014KT006)的研究成果之一。

在上述两类真实数据中，借用菜谱数据分析的研究有：Wagner 等[2]通过常见的复合调味料(Flavor Compounds)中包含的烹饪食材共现关系构造风味网络(Flavor Network)，通过风味网络发现西方烹饪时倾向于使用多种香料形成多种口味混合，比较满足所谓食物配对假设(Food Pairing Hypothesis)；相反，东亚地区烹饪时反对这样。Ahn 等[3]分析多个国家和地区的 56 498 份菜谱，发现西方和东方的饮食差别很大，西方最爱用的 6 种食材是牛奶、黄油、香草、鸡蛋、蔗糖浆和小麦，而东方是酱油、葱、香油、米、大豆和姜。不仅如此，西方的厨师喜欢把有很多共同香料的食材放进同一道菜里面，而东方厨师反对这样。Zhu 等[4]利用从美食杰网站上采集到的 20 个菜系的 8 498 道菜谱和 2 911 种食材，结合菜系所在省市地理位置和气候条件的相似性，分析发现地理上的相近性对于食材使用的影响远远大于气候的相近性，另外针对食材使用矩阵(两个维度分别是菜系和食材)，通过简单的主成分分析找到了云贵菜和香港菜这两个异常菜系。

在上述两类真实数据中，借用用户饮食数据分析的研究有：Ahn 等[5]对大型菜谱网站(ichkoche.at) 的日志进行分析(2012 年 8 月–2013 年 11 月)，发现用户的饮食偏好：用户对菜谱的偏好主要取决于菜的香料；菜谱偏好分布存在的地区差异大于香料偏好分布在地区上的差异；用户工作日的饮食偏好与周末的饮食偏好存在明显不同。Abbar 等[6]根据 Twitter 上 21 万用户的饮食 Tweet 及用户兴趣、地理位置和社交网络数据，分析发现食物的热量与当地肥胖比率存在一定关联，二者皮尔逊相关系数接近 0.77，并基于人口统计变量和 Twitter 上提及的食品名称构建预测地区肥胖和糖尿病人数的模型。

总结上述研究，笔者发现有关饮食社区挖掘的研究尚不多见，而饮食社区挖掘不仅能深层次地挖掘地区用户口味，而且可以发现用户点菜风格，为用户点菜提供指导。因此本文结合菜谱数据与用户饮食评论数据进行饮食社区挖掘的研究。

## 3 研究思路及关键技术

### 3.1 研究思路

结合菜谱数据和微博用户饮食评论数据对饮食社区进行研究，研究思路如图 1 所示：

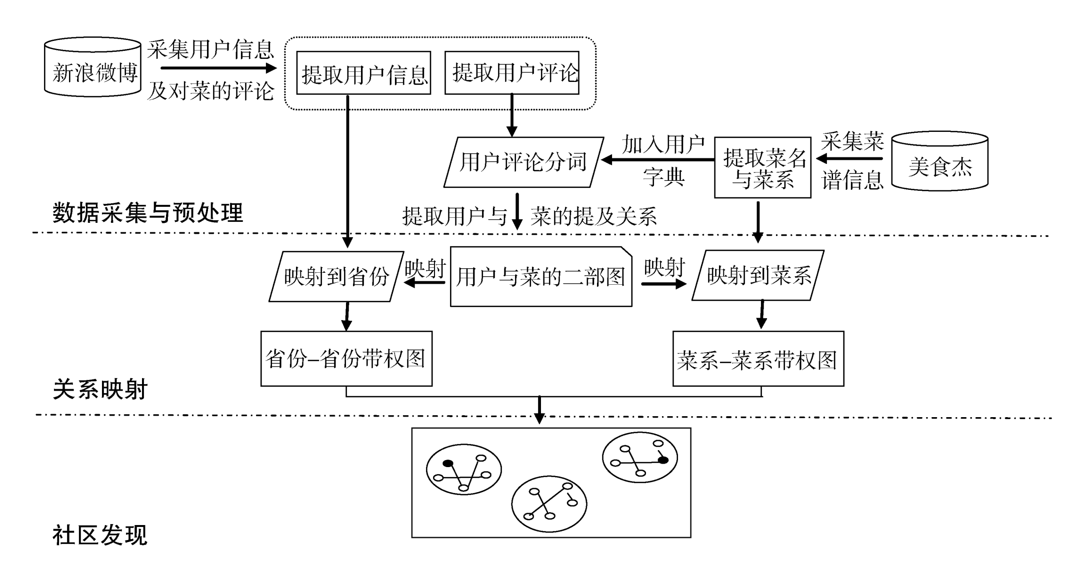



图 1 饮食社区挖掘研究思路

(1) 数据采集与预处理。从“美食杰”网站上采集菜谱名称、菜系等信息，在对“美食杰”网站上采集到的菜名经过简单预处理后，以菜谱名称作为搜索关键词从新浪微博上爬取微博内容及微博用户信息。采集用户微博数据后，将菜谱名称作为用户字典加入分词包，最终完成对用户评论内容的有效分词。

(2) 关系映射。在完成用户评论分词后，根据分词结果提取用户与菜名的“提及”关系(指该菜谱名称出现在用户的微博评论内容中)，根据用户所在省份及菜所属的菜系进行以下两个方面的映射：

①省份–省份关系映射。由于用户有省份信息，如果不同省份用户对同一道菜都有“提及”，可以认为这两个省份之间存在某种共同的联系，如这两地区用户口味相近等，因此可以根据用户与菜之间的“提及”关系，完成省份与省份之间相关关系的映射；

②菜系–菜系关系映射。同样，由于菜有菜系信息，如果同一个用户同时“提及”到不同的菜系，可以认为这两个菜系也存在某种相关性，因此可以根据用户与菜之间的“提及”关系，完成菜系与菜系相关关系的映射。

(3) 社区发现。在完成省份–省份关系映射和菜系–菜系关系映射后，选用合适的社区挖掘算法进行饮食社区挖掘，并完成隐含的省份之间关系和隐含的菜系之间关系的分析与结果可视化。

### 3.2 关键技术

在本文研究中，社区发现技术是主要技术。现有社区发现算法很多，比较经典的方法有Girvan and Newman 的GN分裂算法[7]、Newman等的模块度最大化方法[8]、Shi等的归一化割(Normalized Cut, N-cut)方法[9]、Von Luxburg的基于拉普拉斯矩阵的谱平分方法[10]、LPA算法[11]等，标签传播算法(Label Propagation Algorithm, LPA)[12]是Zhu等于2002年提出的一种基于图的半监督学习方法，其基本思想是用已标记节点的标签信息去预测未标记节点的标签信息。2007年，Raghavan等[11]首次将LPA应用于社区发现，并在Zachary Karate网络[13]、College Football网络[7]等真实基准网上进行测试，结果表明LPA的社区结构检测效果良好。LPA应用于社区发现的步骤如下：

①初始化网络中所有节点的标签，依次为每个节点分配唯一的标签；

②令迭代次数 t=1；

③随机排列网络中的节点，生成序列 X；

④按照序列 X 中的顺序，对 X 中的每个节点 v，使用 $Lv = \arg\max_{l} |N^{l}(v)|$ 更新自身的标签，其中 $N^{l}(v)$ 是拥有 $l$ 标签的 v 的邻居节点集。如果存在多个标签数量最多时，则随机选择其中一个；

⑤如果每个节点具有的标签都是其邻居节点中出现次数最多的标签，算法停止，否则令 t=t+1，转到步骤③。

在进行社区划分时，由Newman等于2004年提出的模块度[8]被用于衡量社区划分质量，模块度计算公式如下：

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j) \qquad (1)$$

其中，$A_{ij}$ 是网络图的邻接矩阵，m 是图中边的总数，$P_{ij}$ 表示在空模型中顶点 i 和 j 间边的期望值，如果顶点 i 和 j 在同一个社区，$\delta(C_i, C_j) = 1$，否则为 0。而且公式(1)在邻接矩阵和节点度数上作相应计算后亦可以使用在带权图上[14]，因此本文使用该公式作为社区划分评判依据。

## 4 实验及实验结果

### 4.1 实验数据集

从国内知名网站“美食杰”①上采集 20 个菜系的菜谱，菜谱信息主要包括菜名、菜类别、菜系、菜的主料、辅料及菜的做法等，对 Zhu 等[4]的菜名②过滤后，得到包含“川菜”、“东北菜”、“港台菜”、“其他菜”、“湖北菜”、“沪菜”、“徽菜”、“江西菜”、“京菜”、“鲁菜”、“闽菜”、“清真菜”、“山西菜”、“苏菜”、“西北菜”、“湘菜”、“豫菜”、“粤菜”、“云贵菜”、“浙菜”等 20 个菜系的 5 156 份有用菜谱，各个菜系的数目如表 1 所示：

表 1 各个菜系下菜谱数目统计

| 菜系 | 菜谱数 | 菜系 | 菜谱数 |
|---|---|---|---|
| 川菜 | 743 | 东北菜 | 227 |
| 鲁菜 | 598 | 徽菜 | 143 |
| 粤菜 | 491 | 西北菜 | 119 |
| 沪菜 | 454 | 湖北菜 | 109 |
| 京菜 | 380 | 豫菜 | 104 |
| 湘菜 | 370 | 港台菜 | 89 |
| 闽菜 | 286 | 江西菜 | 87 |
| 清真菜 | 283 | 山西菜 | 85 |
| 浙菜 | 271 | 云贵菜 | 51 |
| 苏菜 | 242 | 其他菜 | 24 |

根据表 1 中的菜名在微博上搜集并采集相关信息，最终累计采集到来自 36 个不同地区(包括用户填写的“其他”和“海外”的地区在内)的共计 3 980 597 个用户的 8 746 931 条微博信息。其中，各条微博信息包括用户发布微博时间、微博内容、用户省份、用户和性别等。在采集到相关微博后，先将保留下来的菜名作为用户字典加入结巴分词包(jieba)，对用户饮食微博内

①http://www.meishij.net.

②https://github.com/zhuyuxiao/Chinese-cuisine.

容进行分词切分，在用户微博内容的分词结果中只提取包含在菜谱中菜名内容后，得到 2 269 763 个用户与菜之间的“提及”关系，作为本文社区发现的主要研究对象，其中，统计“提及”关系中涉及到的不同省份地区的用户人数如表 2 所示。

### 4.2　关系映射

为了分析省份地区用户口味及菜系在用户点餐中的情况，对这种“提及”关系在地区维度和菜系维度上分别进行映射，得到地区–地区带权图和菜系–菜系带权图。由于用户与菜之间“提及”关系的规模较大，得到的带权图规模有限，但是边的权重很大，以各个地区及各个菜系为例，给出最亲密(即权重最大)的节点，如表 3 和表 4 所示。

表 2　不同地区的用户数统计

| 地区 | 用户数 | 地区 | 用户数 | 地区 | 用户数 |
|---|---|---|---|---|---|
| 广东 | 334 681 | 河南 | 51 404 | 山西 | 14 414 |
| 北京 | 263 762 | 辽宁 | 49 950 | 新疆 | 13 433 |
| 上海 | 209 098 | 湖南 | 43 693 | 贵州 | 12 757 |
| 江苏 | 146 130 | 陕西 | 38 746 | 香港 | 11 664 |
| 海外 | 138 643 | 安徽 | 34 427 | 海南 | 11 477 |
| 浙江 | 129 848 | 天津 | 34 089 | 内蒙古 | 8 923 |
| 四川 | 128 761 | 河北 | 29 954 | 甘肃 | 7 383 |
| 其他 | 128 597 | 广西 | 29 791 | 台湾 | 6 450 |
| 福建 | 98 092 | 江西 | 24 985 | 西藏 | 3 453 |
| 湖北 | 71 562 | 黑龙江 | 21 724 | 宁夏 | 2 430 |
| 山东 | 70 590 | 云南 | 19 846 | 青海 | 2 160 |
| 重庆 | 57 772 | 吉林 | 14 852 | 澳门 | 2 117 |

表 3　地区映射图中各个地区最紧密的地区及权重

| 地区 | 最紧密地区(权值) | 地区 | 最紧密地区(权值) | 地区 | 最紧密地区(权值) |
|---|---|---|---|---|---|
| 青海 | 北京(4 286) | 宁夏 | 海外(4 778) | 海外 | 浙江(200 138) |
| 辽宁 | 北京(96 520) | 湖南 | 其他(71 374) | 陕西 | 浙江(69 376) |
| 贵州 | 北京(25 044) | 台湾 | 海外(12 466) | 山西 | 浙江(25 100) |
| 北京 | 广东(385 844) | 河北 | 其他(58 526) | 新疆 | 浙江(22 374) |
| 广西 | 广东(59 064) | 西藏 | 江苏(6 394) | 四川 | 浙江(141 378) |
| 澳门 | 广东(4 206) | 其他 | 海外(224 654) | 重庆 | 浙江(87 692) |
| 广东 | 上海(338 516) | 吉林 | 海外(28 594) | 湖北 | 浙江(119 244) |
| 上海 | 江苏(252 126) | 黑龙江 | 海外(41 092) | 江苏 | 浙江(216 320) |
| 海南 | 其他(20 254) | 福建 | 浙江(148 500) | 河南 | 浙江(91 806) |
| 甘肃 | 其他(14 630) | 天津 | 海外(65 058) | 浙江 | 浙江(129 604) |
| 山东 | 江苏(124 424) | 内蒙古 | 海外(17 360) | 香港 | 云南(18 250) |
| 江西 | 其他(46 206) | 安徽 | 江苏(66 604) | 云南 | 云南(19 817) |

表 4　菜系映射图中各个菜系最紧密的菜系及权重

| 菜系 | 最紧密的菜系 | 权重 | 最紧密的菜系 | 菜系 | 权重 |
|---|---|---|---|---|---|
| 粤菜 | 粤菜 | 42 049 | 鲁菜 | 鲁菜 | 61 144 |
| 京菜 | 川菜 | 104 974 | 云贵菜 | 云贵菜 | 4 299 |
| 浙菜 | 浙菜 | 22 904 | 湘菜 | 湘菜 | 18 731 |
| 川菜 | 川菜 | 373 359 | 西北菜 | 西北菜 | 21 559 |
| 湖北菜 | 湖北菜 | 21 457 | 其他菜 | 其他菜 | 4 621 |
| 闽菜 | 闽菜 | 53 853 | 豫菜 | 豫菜 | 9 717 |
| 苏菜 | 苏菜 | 40 233 | 山西菜 | 山西菜 | 11 191 |
| 徽菜 | 徽菜 | 40 716 | 清真菜 | 清真菜 | 23 195 |
| 港台菜 | 港台菜 | 18 067 | 沪菜 | 沪菜 | 103 096 |
| 江西菜 | 江西菜 | 16 237 | | | |

通过表 3 可以看出，除了“其他”、“海外”这两个未知地区外，“浙江”、“江苏”、“广东”、“北京”地区与其余省份联系极为紧密，这可能因为这些地区人数较多的缘故。进一步对菜系的紧密程度进行分析，如表 4 所示。可以看出在这 20 个菜系中，有 18 个菜系与自身“共现”(此处的“共现”指被同一个用户提及)频次最高，只有“京菜”与“川菜”共现最高，这说明“川菜”很受用户喜爱。

### 4.3　饮食社区挖掘结果

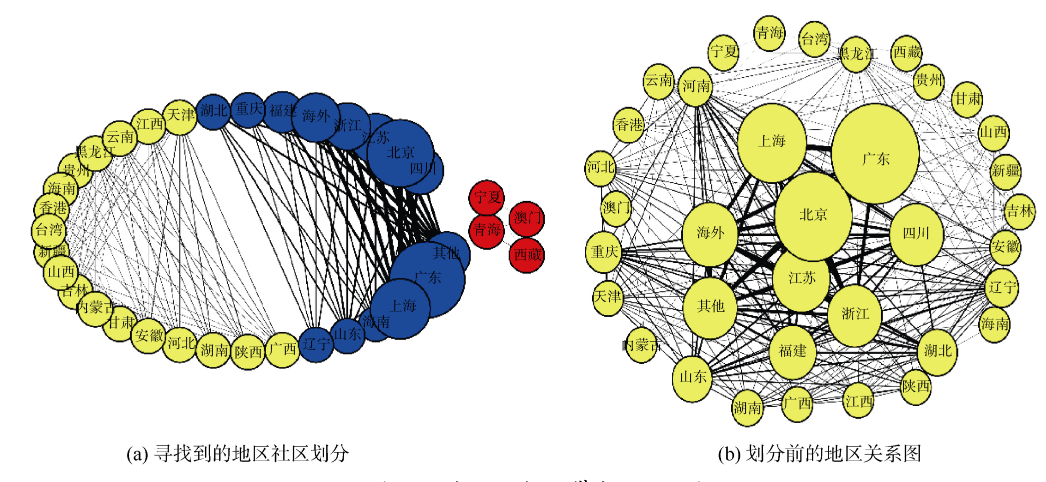
在完成映射后，笔者试图对地区完全带权图和菜系完全带权图进行社区发现，但发现这个完全带权

图无法进行社区划分，经过多次迭代后，模块度均为0，这可能是因为完全带权图上并不存在通常意义上的社区，即每个社区内部节点间的连接相对非常紧密，但是各个社区之间的连接却相对比较稀疏[8]。为此，先对这两个完全带权图进行断边处理，让一些节点连接边断开。通常，断边操作发生在网络病毒传播的控制过程，用于有效抑制病毒的传播[15]，本文应用断边处理主要是因为所得的网络图是完全图不适合进行社区划分。根据前面的分析结果，发现本文所得的完全图边上存在权重(反映不同节点之间的共现次数)，节点也有权重(反映节点自身共现的次数)，而根据吴亮等[16]的研究：只有节点权重值至少是接近的两节点之间，才有可能出现同步或者说至少是两节点行为之间存在关联的现象。因此，如果节点权重差距悬殊，本文认为这两个节点很难划分进一个社区(社区通常是由功能相近或性质相似的网络节点组成)。为此，笔者切断节点权重悬殊较大的节点之间的边，经过多次实验测试，最终选择节点间的边权作为判别节点权重差距悬殊(断边)的依据，如公式(2)所示：

$$|W_p - W_q| > W_{Epq} \quad (2)$$

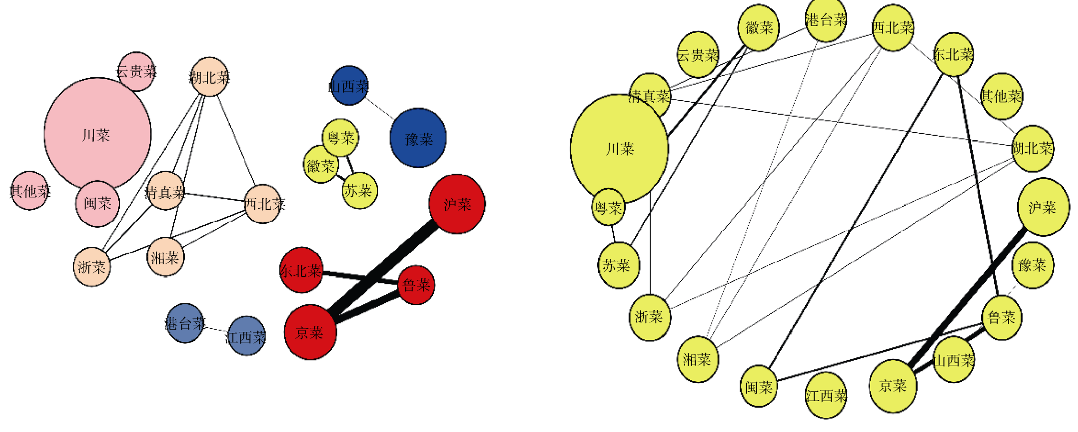
其中，$W_p$、$W_q$ 分别代表节点 p 和节点 q 的权重，$W_{Epq}$ 是节点 p 和节点 q 之间边 $E_{pq}$ 的权重。

公式(2)断边后，在包含自连接的含有 666 条边的地区完全带权图与 210 条边的菜系完全带权图边数分别下降为 261 和 40。对断边后的地区带权图与菜系带权图利用 LPA 进行社区发现，分别执行 100 次，取模块度最大的结果见图 2(a)(Q=0.370)和图 3(a)(Q=0.695)，而各个图划分前的结果见图 2(b)与图 3(b)(图 2 和图 3 中节点大小反映节点与自身共现权重大小，边粗细反映节点与节点间共现权重大小)。




(a) 寻找到的地区社区划分　　(b) 划分前的地区关系图

图 2　地区–地区带权网络图




(a) 寻找到的地区菜系划分　　(b) 划分前的菜系关系图

图 3　菜系–菜系带权网络图

由图 2 可以看出，最终被划分为三个社区(邻居颜色相同的为一个社区):

(1) 社区 1: 青海、澳门、宁夏、西藏;

(2) 社区 2: 贵州、广西、云南、甘肃、江西、海南、天津、河北、吉林、黑龙江、湖南、台湾、内蒙古、安徽、陕西、山西、新疆、香港;

(3) 社区 3: 辽宁、北京、广东、上海、山东、福建、其他、海外、四川、重庆、湖北、江苏、河南、浙江。

通过划分结果可以直观地看出:

(1) 社区 1 主要由自治区组成;

(2) 社区 3 主要由我国比较发达地区组成，如北京、上海、广东、江苏、浙江、福建，说明这些地区用户讨论的菜(即用户口味)比较接近。

为了进一步比较各个社区内用户口味，对各个社区内省份用户"提及"的菜名及"提及"次数进行统计，得到次数居前的 Top7 菜单见表 5 左侧，同时对这些地区特色菜进行提取(通过各个社区内前 100 的菜名的差集运算得出)，结果见表 5 右侧。

通过表 5 左侧数据，可以看出各个社区内地区的对"烤肉"、"酸辣粉"、"毛血旺"等菜都有共同的爱好；通过右侧数据，可以看出各个社区用户口味的差异菜单，结合菜谱数据中这些菜的口味后，分析发现社区 1 内、社区 2 内及社区 3 内的地区特有菜分别属于"其他口味"、"鲜咸味"、"香辣味"。

表 5　省份地区用户"提及"菜统计

| 社区 1 内地区 Top7 的菜 | 社区 2 内地区 Top7 的菜 | 社区 3 内地区 Top7 的菜 | 社区 1 内地区特有的菜 | 社区 2 内地区特有的菜 | 社区 3 内地区特有的菜 |
|---|---|---|---|---|---|
| 抓炒鱼条 | 烤肉 | 清汤鱼圆 | 盐爆鱿鱼卷 | 西红柿炒肉片 | 水煮活鱼 |
| 云片鹿角菜 | 水煮肉片 | 毛血旺 | 灯影牛肉 | 风霜雪叶 | 瓜仁西葫芦 |
| 烤肉 | 清汤鱼圆 | 烤肉 | 焦盐子芋 | 蛋酥花仁 | 六月鲜 |
| 清汤鱼圆 | 毛血旺 | 春笋白拌鸡 | 青椒素肉丝 | 冬瓜咸肉 | 蛋烧麦 |
| 麻辣烫 | 酸辣粉 | 水煮肉片 | 蚝油鸭脚 | 葫芦鸡 | 莲蓬豆腐 |
| 酸辣粉 | 粉蒸肉 | 酸辣粉 | 干炸肝花 | 缠丝鸡饼 | 香辣猪蹄 |
| 腌鲜鳜鱼 | 腌鲜鳜鱼 | 炸烹螃蟹 | 雪银虾饼 | 古老肉 | 干锅排骨 |

由图 3 可以看出，根据菜系共现关系，菜系被划分成以下 9 个社区:

(1) 社区 1: 湘菜、湖北菜、西北菜、清真菜、浙菜;

(2) 社区 2: 京菜、沪菜、鲁菜、东北菜;

(3) 社区 3: 粤菜、徽菜、苏菜;

(4) 社区 4: 山西菜、豫菜;

(5) 社区 5: 港台菜、江西菜;

(6) 社区 6: 闽菜;

(7) 社区 7: 川菜;

(8) 社区 8: 云贵菜;

(9) 社区 9: 其他菜。

通过划分结果可以直观地看出:

(1) "川菜"、"闽菜"、"云贵菜"、"其他菜"自成一体。其中"云贵菜"与"其他菜"被分成独立的社区，与文献[4]中发现的结果相一致;

(2) 社区 1、社区 3、社区 4 体现了菜系的地理位置近邻性。这是因为: 根据 Zhu 等[4]的研究: "西北菜"起源于"陕西、甘肃、青海、宁夏"，"清真菜"起源地"新疆"，"豫菜"起源于"山东"，因此这几个社区体现了菜系具有一定的地理位置近邻性。

为了进一步分析"川菜"、"云贵菜"为何被划分成独立菜系及解释社区 2、社区 5，基于 Zhu 等认辅料①为粒度，进一步分析各个菜的辅料比例(计算方法为: 菜系下辅料总数除以菜系下菜的数目)，结果如表 6 所示。

通过表 6 可以发现，"川菜"在原料方面确实比较偏"辣"(除了常见的油盐酱醋等辅料外，"花椒"、"胡椒"、"辣椒"、"豆瓣酱"等也较多地被使用)，"云贵菜"比较与众不同，很少使用"味精"调料，而较多地使用"虾"，因此这两个菜系也被划分成独立菜系；另外，还可以看出社区 2 内"京菜"、"沪菜"、"鲁菜"、"东

①https://github.com/zhuyuxiao/Chinese-cuisine.

表 6　地区菜系辅料使用比例统计

| 川菜 | 云贵菜 | 京菜 | 沪菜 | 鲁菜 | 东北菜 | 江西菜 | 港台菜 |
|---|---|---|---|---|---|---|---|
| 盐(0.87) | 盐(0.75) | 盐(0.83) | 盐(0.8) | 盐(0.89) | 盐(0.83) | 盐(0.84) | 盐(0.77) |
| 姜(0.69) | 料酒(0.5) | 香葱(0.69) | 味精(0.7) | 香葱(0.7) | 香葱(0.77) | 味精(0.69) | 白糖(0.51) |
| 香葱(0.66) | 姜(0.5) | 姜(0.68) | 白糖(0.67) | 姜(0.63) | 姜(0.67) | 姜(0.61) | 香葱(0.47) |
| 味精(0.63) | **虾(0.25)** | 味精(0.68) | 姜(0.59) | 味精(0.6) | 料酒(0.44) | 香葱(0.58) | 酱油(0.44) |
| 料酒(0.47) | 冰糖(0.25) | 白糖(0.51) | 香葱(0.56) | 淀粉(0.46) | 酱油(0.44) | 料酒(0.46) | 蒜(0.37) |
| 白糖(0.47) | 醋(0.25) | 料酒(0.45) | 酱油(0.38) | 料酒(0.4) | 白糖(0.41) | 酱油(0.42) | 胡椒(0.26) |
| 酱油(0.37) | 猪油(0.25) | 酱油(0.42) | 淀粉(0.37) | 酱油(0.4) | 味精(0.41) | 香油(0.4) | 味精(0.26) |
| 蒜(0.33) | 胡椒(0.25) | 淀粉(0.41) | 香油(0.29) | 白糖(0.39) | 花椒(0.4) | 猪油(0.36) | 香油(0.23) |
| **胡椒(0.33)** | 植物油(0.25) | 香油(0.3) | 花生油(0.27) | 鸡蛋(0.34) | 蒜(0.36) | 淀粉(0.27) | 姜(0.21) |
| 淀粉(0.31) | 低筋面粉(0.25) | 鸡蛋(0.3) | 黄酒(0.27) | 香油(0.3) | 淀粉(0.28) | 辣椒(0.27) | 鸡蛋(0.19) |
| **花椒(0.31)** | 白胡椒(0.25) | 花生油(0.23) | 猪油(0.23) | 蒜(0.24) | 鸡蛋(0.25) | 鸡蛋(0.27) | **高汤(0.14)** |
| 香油(0.25) | 香葱(0.25) | 蒜(0.21) | 料酒(0.21) | 醋(0.23) | 醋(0.23) | 胡椒(0.25) | 辣椒(0.14) |
| 鸡蛋(0.23) | 香芹(0.25) | 花椒(0.21) | 胡椒(0.18) | 花椒(0.22) | 香油(0.17) | 香菇(0.24) | 米酒(0.14) |
| 醋(0.22) | 苏打粉(0.25) | 醋(0.2) | 竹笋(0.17) | 植物油(0.2) | 植物油(0.17) | 蒜(0.24) | 太白粉(0.12) |
| **辣椒(0.2)** | 带鱼(0.25) | 黄酒(0.18) | 鸡蛋(0.16) | 香菜(0.17) | 香菜(0.15) | 猪肉(0.22) | 虾米(0.12) |
| **豆瓣酱(0.19)** | 小麦面粉(0.25) | 猪油(0.17) | 醋(0.14) | 胡椒(0.17) | 猪肉(0.12) | 白糖(0.2) | 麻油(0.12) |

北菜”口味比较接近，尤其“东北菜”与“鲁菜”，这与“东北菜”是“鲁菜”的一个分支也相一致；至于社区 5 内“港台菜”、“江西菜”被划分成一个社区，从菜的辅料上较难解释，但通过“港台菜”的原料可以看出，“高汤”成为这个菜系的一个主要辅料。

## 5　结　语

本文结合菜谱信息和微博用户评论内容进行饮食社区研究，结果发现：

(1) 省份地区被划分成“鲜咸味”、“香辣味”、“其他口味”这三个口味地区；

(2) “川菜”、“云贵菜”因辅料独特很少与其他菜系被一起点餐，“京菜”、“沪菜”、“鲁菜”、“东北菜”常被一起点餐；

(3) 地区菜系体之间存在一定程度的地理位置近邻性。

相对饮食文化文献、史料等统计整理的计量方法来说，本文基于真实数据集上所得结论会更有说服力，但也存在一些不足，即所获得的微博用户人数受地区差异性影响大，发达地区人数与边缘地区人数悬殊太大，这可能对本文所得结论有一定的影响，希望未来能找到更好的办法克服地区人口差异性以进行更好的分析。

在完成饮食社区挖掘后，粗略地给出了各个省份地区划分与地区菜系划分，并没有深入地挖掘各个省份社区内用户口味及地区菜系点餐背后隐含的潜在关联。除此之外，本文只考虑了“提及”关系，没有对这种“提及”关系的正负情感进行分析，这些将是下一步研究工作。

## 参考文献：


[1] 陈国林. 饮食文化学：研究概述与学科距离[J]. 四川烹饪高等专科学校学报, 2013(2): 4-7. (Chen Guolin. Study of Food Culture: An Overiew and Its Constraints [J]. Journal of Sichuan Higher Institute of Cuisine, 2013(2): 4-7.)

[2] Wagner C, Singer P, Strohmaier M. The Nature and Evolution of Online Food Preferences [J]. EPJ Data Science, 2014, 3(1): Article No. 38.

[3] Ahn Y Y, Ahnert S. The Flavor Network [J]. Leonardo, 2013, 46(3): 272-273.

[4] Zhu Y X, Huang J, Zhang Z K, et al. Geography and Similarity of Regional Cuisines in China [J]. PLoS ONE, 2013, 8(11): e79161.

[5] Ahn Y Y, Ahnert S E, Bagrow J P, et al. Flavor Network and the Principles of Food Pairing [OL]. arXiv: 1111.6074.

[6] Abbar S, Mejova Y, Weber I. You Tweet What You Eat: Studying Food Consumption Through Twitter [OL]. arXiv Preprint, 2014. arXiv: 14124361.

[7] Girvan M, Newman M E. Community Structure in Social and Biological Networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.
[8] Newman M E, Girvan M. Finding and Evaluating Community Structure in Networks [J]. Physical Review E, 2004, 69(2): 026113.
[9] Shi J, Malik J. Normalized Cuts and Image Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.
[10] Von Luxburg U. A Tutorial on Spectral Clustering [J]. Statistics and Computing, 2007, 17(4): 395-416.
[11] Raghavan U N, Albert R, Kumara S. Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks [J]. Physical Review E, 2007, 76(3): 036106.
[12] Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation[R]. Carnegie Mellon University, 2002. http://discovery.ucl.ac.uk/id/eprint/185718.
[13] Zachary W W. An Information Flow Model for Conflict and Fission in Small Groups [J]. Journal of Anthropological Research, 1977, 33(4): 452-473.
[14] Newman M E. Analysis of Weighted Networks [J]. Physical Review E, 2004, 70(5): 056131.
[15] 宋玉蓉，蒋国平，徐加刚. 一种基于元胞自动机的自适应网络病毒传播模型[J]. 物理学报，2011，60(12): 110-119. (Song Yurong, Jiang Guoping, Xu Jiagang. An Epidemic Spreading Model in Adaptive Networks Based on Cellular Automata [J]. Acta Physica Sinica, 2011, 60(12): 110-119.)
[16] 吴亮，朱士群. 网络中的节点权重及其物理意义[C]. 见：第十二届全国量子光学学术会议论文摘要集. 2006. (Wu Liang, Zhu Shiqun. Node Weights and Its Physical Significance in Netwroks [C]. In: Proceedings of the 12th National Symposium on Quantum Optics. 2006.)

## 作者贡献声明:

吴小兰：文献调研与整理，进行实验，论文起草；
章成志：提出研究思路，讨论研究方案，论文最终版本修订。

## 利益冲突声明:

所有作者声明不存在利益冲突关系。

## 支撑数据:

支撑数据由作者自存储, E-mail: wuxiaolananhui@163.com。

[1] 吴小兰，章成志. userid&province.txt. 用户 id 及其提及的菜名与菜系.
[2] 吴小兰，章成志. userid&dietname.txt. 用户 id 及用户所属省份.
[3] 吴小兰，章成志. province_network.txt. 省份维度上映射后的省份网络.
[4] 吴小兰，章成志. diet_network.txt. 菜系维度上映射后的菜系网络.
[5] 吴小兰，章成志. diet_network(cut_edge).txt. 进行断边后的菜系网络.
[6] 吴小兰，章成志. province_network(cut_edge).txt. 进行断边后的省份网络.

收稿日期: 2016-03-17
收修改稿日期: 2016-03-30

# Analyzing Food Community with Recipes and Weibo User Reviews

Wu Xiaolan[1, 2] Zhang Chengzhi[2, 3]
[1](School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China)
[2](Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China)
[3](Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210093, China)

**Abstract:** [**Objective**] This study examines the structure of online food community with the help of large-scale real world data. [**Methods**] First, we collected recipes from meishij.net (a popular food network online) and user reviews from Sina Weibo (micro-blog) respectively. Second, we identified the Weibo users who mentioned recipes from meishij.net and mapped them to provinces and cuisines coordinate systems. Finally, we used community discovery algorithm to analyze the food community's structure. [**Results**] The province and cuisines networks showed clear community structures. [**Limitations**] Demographic disparity might pose some effects to the conclusions. [**Conclusions**] The tastes of consumers from different provinces could be classified as "freshly salty", "hot and spicy", as well as "others". "Sichuan" or "Yungui" dishes are rarely ordered together, while "Jing", "Hu", "Lu" and "Dongbei" dishes are often ordered along with each other. Besides, the regional cuisines have some geographical proximity among themselves.

**Keywords:** Food culture Regional cuisines Food community Web information organization



## CCC 进一步增强 RightFind 内容工作流解决方案

版权结算中心有限公司(Copyright Clearance Center, CCC)是一家致力于创造全球许可和版权内容解决方案的公司，其于近日公布了增强其基于云服务的 RightFind 内容的工作流解决方案。

RightFind 为用户提供数千种期刊即时、便捷的访问，同时还能帮助管理者优化在采购和管理内容上的支出。RightFind 7.0 主要包含以下三方面的功能增强:

(1) 一个升级的用户界面，以简化工作流程，并使得查找内容变得更加容易;

(2) 引入 CrossRef 数据以加速 RightFind 上新近出版文献被引信息的揭示;

(3) 增加两个新的 API，允许用户从 RightFind 库中提取信息，从而使得他们能够在其他应用程序中搜寻 RightFind 的内容。

CCC 公司产品和服务总监 Lauren Tulloch 说: "我们的客户追求与信息之间的无缝衔接。我们正在加强平台建设，以提升 RightFind 的用户体验，同时使得其他应用从 RightFind 中提取数据并进行利用变得更加容易。"

作为 RightFind 内容工作流程解决方案组件的一部分，CCC 还提供 RightFind XML 以供文本挖掘。生命科学领域的科研人员可以为来自 6 000 多份同行评议期刊的 500 多万篇文章创建带全文的 XML 文件，并在第三方文本挖掘软件中进行遵循版权的使用。

(编译自: http://www.copyright.com/copyright-clearance-center-announces-latest-enhancements-to-rightfind-content-workflow-solution/)

(本刊讯)